BAB II
TINJAUAN PUSTAKA

2.1 Konsep Dasar Pengertian

2.1.1 Data Mining

Banyak pihak yang telah mendefiniskan *data mining*. Berikut beberapa definisi *data mining*:

"*Data mining* merupakan analisis dari sekumpulan data yang diamati (sangat besar) untuk menemukan hubungan yang tidak terduga dan merangkum data dengan cara yang baru yang dapat dipahami dan bergunak bagi pemilik data."(Hand, et al)[3].

Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang tersimpan di dalam *database* besar [4].

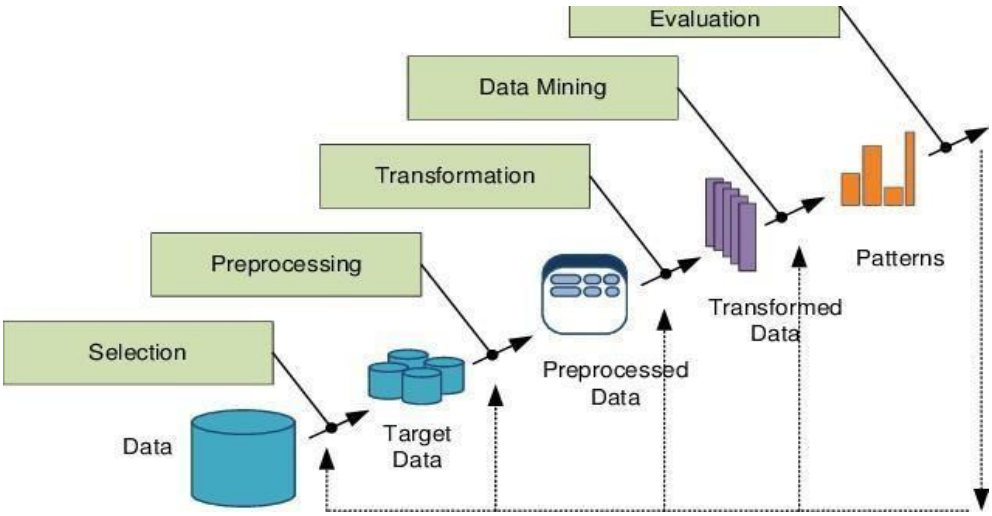
2.1.2 Prediksi/Forecasting

Prediksi/peramalan *(forecasting)* merupakan suatu kegiatan meramalkan penjualan dimasa mendatang berarti menentukaan perkiraan besarnya volume penjualan, bahkan menentukan potensi penjualan dan luas pasar yang dikuasai dimasa mendatang. Selain itu membantu perusahaan dalam melakukan perencanaan penyediaan stok, karena prediksi ini memberikan *output* terbaik bagi perusahaan sehingga dapat meminimalisir kesalahan perencanaan dapat ditekankan seminimal mungkin [5].

2.1.3 *Knowledge Discovery in Database*

*Knowledge Discovery in Database (*KDD) sebagai proses dari menggunakan metode data mining untuk mencari informasi- informasi

yang berharga, pola yang ada di dalam data, yang melibatkan algoritma untuk mengidentifikasi polapada data (Fayyed EtAl 1996) [6]. Berikut tahapan proses KDD dapat dilihat pada gambar 2.1



Gambar 2. 1 Tahapan dalam KDD

Tahapan Proses KDD terdiri dari:

1. *Data Selection*

a. Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan.

b. Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing/ Cleaning*

a. Pemprosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan.

b. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD.

c. Proses *cleansing* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

d. Dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformation

a. Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantungkepada goal yang ingin dicapai.

b. Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

a. Pemilihan tugas data mining; pemilihan goal dari proses KDD misalnya klasifikasi, regresi, clustering, dll.

b. Pemilihan algoritma data mining untuk pencarian (searching).

c. Proses Data mining yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan prosesKDD secara keseluruhan.

5. Interpretation/ Evaluation

a. Penerjemahan pola-pola yang dihasilkan dari data mining.

b. Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah mimengerti oleh pihak yang berkepentingan.

c. Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

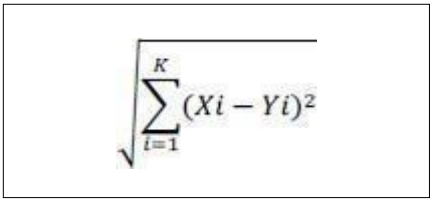
2.1.4 K-Nearst Neighbour

Metode *K-Nearest Neighbour* (KNN) merupakan salah satu metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Tujuannya adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan data training. Klasifikasi dilakukan tanpa menggunakan model tetapi hanya berdasarkan memori. Algoritma *K-Nearest Neighbour* menggunakan klasifikasi ketetanggaan sebagai prediksi terhadap data baru [6].

Langkah-langkah untuk menghitung algoritma K-NN:

1. Menentukan nilai k.
2. Menghitung kuadrat jarak *euclid* (*query instance*) masing-masing objek terhadap *training data* yang diberikan.
3. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yangmempunyai jarak *Euclidean* terkecil.
4. Mengumpulkan label *class* Y (klasifikasi *Nearest Neighbor*).
5. Dengan menggunakan kategori *Nearest Neighbor* yang paling mayoritas maka dapat dipredeksikan nilai *query instance* yang telah dihitung.

Pada penelitian ini penulis



Gambar 2. 2 Rumus Perhitungan jarak Euclidiean

Nilai Xi merupakan nilai yang ada pada data *training*, sedangkan nilai Yi merupakan nilai yang ada pada data *testing*. Nilai K merupakan dimensi atribut.

2.1.5 Rapid Miner

RapidMiner (YALE) adalah perangkat lunak *open source* untuk *knowledge discovery* dan *data mining*. *Rapidminer* memiliki kurang lebih 400 prosedur (operator) *data mining* termasuk operator untuk masukan, *output*, data *preprocessing* dan visualisasi (Sulianta, dkk 2010:101) [6].

Beberapa fitur dari *rapidminer*, antara lain:

1. Berlisensi gratis (*open source*).
2. Multiplatform karena diprogram dalam bahasa Java.
3. Internal data berbasis XML sehingga memudahkan pertukaran data eksperimen.
4. Dilengkapi dengan *scripting language* untuk otomatisasi eksperimen.
5. Memiliki GUI (*Graphical User Interface*), command line mode (batch mode), dan Java API yang dapat dipanggil dari program lain.
6. Dapat dikembangkan dengan menambahkan plugin dan ekstension.

Fasilitas *plotting* untuk visualisasi data multidimensi dan model.

Gambar 2. 3 Aplikasi *Rapid Miner*

2.2 Penelitian Terkait

Dalam penelitian ini, penulis mengacu pada beberapa referensi sejenis, diantara lain:

Tabel 2. 1 Penelitian Terdahulu

No	Peneliti	Judul	Metode	Hasil
1	Sri Puspita Dewi, Nurwati, Elly Rahayu	Penerapan Data Mining Untuk Prediksi Penjualan Porduk Terlaris Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Hasil penelitian yang telah dilakukan dapat disimpulkan bahwa sistem aplikasi yang dibuat dapat membantu dalam menentukan prediksi penjualan produk terlaris pada UD Andar
2	Inna Alvi Nikmatun, Indra Waspada	Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Hasil dari penelitian ini yaitu telah dilakukan 6 skenario percobaan untuk dapat melakukan klasifikasi masa studi yang dimana telah diperoleh nilai akurasi tertinggi pada scenario yang menggunakan atribut mata kuliah pilihan yaitu 75,95%.

Tabel 2. 2 Penelitian Terdahulu (lannjutan)

3	Ike Yolanda, Hasanul Fahmi	Penerapan Data Mining Uuntuk Prediksi Penjualan Produk Roti Terlaris Pada PT Nippon Indosari Tbk Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Disimpulkan bahwa hasil penelitian ini adalah perancangan aplikasi yang berguna bagi perusahaan untuk dapat menentukan produk roti yang terlaris sehingga tidak adnya kerugian yang dialami oleh pihak perusahaan.
4	Yulia Rizki Amalia	Penerapan Data Mining Untuk Prediksi Penjualan Produk Elektronik Terlaris Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Berdasarkan penelitian yang telah dilakukan makan hasil yang didapatkan yaitu Prediksi yang akurat yang telah diteliti menggunakan metode *k-nearest neigbor* yang menghasilakn 6 jenis produk yang akan laris.
5	Aisha Alfani W.P.R, Fahrur Rozi, Farid Sukmana	Prediksi Penjualan Produk Unilever Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Hasil penelitian adalah bahwa menemukan hasil prediksi yang tepat berdasarkan hasil akurasi tertinggi dan terendah yaitu hasil akurasi tertinggi sebesar 86,66% sedangkan terendah yaitu 40%.[8]